Details of Run 1 "top likelihood" estimation for top quark mass analysis in the lepton+jets channel

Mark Strovink
University of California, Berkeley

19 Apr 02

I. INTRODUCTION

This note is intended to provide additional details to DØ physicists planning to measure the top quark mass in the lepton+jets channel in Run 2, who wish to understand how the data for Run 1 top mass analysis were selected, especially how the "top likelihood" for each event was estimated. Their primary references should be the original Run 1 analysis note [1] and the eventual long PRD article [2]. In the context of these primary references, this note is a supplement to [1].

This note has two scopes. First, in Section II we define certain variables that discriminate top from background events without significantly biasing the apparent top mass of the sample. Similar information is provided in [1]; the definitions are included here for completeness and precision. These variables underlay both the likelihood and neural network (NN) determinations by DØ of the top mass in this channel.

Distributions in these variables were combined into a likelihood discriminant $\mathcal{D}_{\rm LB}$ for Run 1 top mass analysis. This was used in one of the two methods that were employed; the other method used a discriminant $\mathcal{D}_{\rm NN}$ taken from the output of a neural network that was based on the same variables. Either discriminant ranged from 0 (background-like) to 1 (top signal-like). Originally, in either branch of the analysis, DØ intended merely to place a cut on the appropriate $\mathcal{D}$; soon it was realized that the error could be reduced by fitting the data to two-dimensional templates in $\mathcal{D}$ *vs.* reconstructed top mass.

The second scope of this note (Section IV) describes how and why the likelihood discriminant $\mathcal{D}_{\rm LB}$ was defined as it was.

The "top likelihood" work was carried out during 1995-1996, collaboratively by M.S. and S. Protopopescu for the first scope, and by M.S. for the second. E. Varnes provided DØ data and Monte Carlo events in a compact format suitable for many iterations of the work described here.

II. DEFINITION OF KEY VARIABLES

Precut sample

The data for all top quark mass analysis in the lepton+jets channel arose from a "precut sample" satisfying certain basic (and two not-so-basic) cuts. Jets, corrected for energy scale and for the energy of any associated tag muon, were required to have $E_T > 15$ GeV and $|\eta| < 2$. Isolated electrons satisfied the same η cut, but isolated muons were restricted to $|\eta| < 1.7$ due to limitations of the forward muon system; both types of lepton satisfied the stiffer cut $E_T > 20$ GeV. Likewise, missing E_T was required to exceed 20 GeV. (In addition, if measured only in the calorimeter, missing E_T was required to exceed 20 (25) GeV for μ+jets (e+jets) events.) Though 3-jet events were accepted for top *cross section* analysis (if one of the jets was b tagged by a non-isolated muon), events for top *mass* analysis were required to have at least 4 jets, in addition to an isolated lepton and missing E_T.

These basic cuts may seem straightforward, but, with data sparse and multivariate analysis methods now widely accepted, one could ask whether, in retrospect, the selection could have been made more efficient. Though electrons were identified by the best means we had – a 5-variable likelihood (or 4-variable when TRD information was unavailable) – the lack of E/p and preshower discrimination caused fake electrons to present a much larger background than is expected for Run 2. Electron fakes had to be modeled by using data (which were sparse) and by assuming that different stages of electron rejection fac-

torize completely. To proceed with confidence, one needed to bring the fakes down to a minor (∼25%) fraction of the more calculable W+jets background. This goal motivated the electron E_T and η cut and the missing E_T cut (and, by symmetry, the muon E_T cut). If we had available an "electron signal-to-noise" in which we had full confidence, it would have been more efficient to cut on that instead.

The jet η cut was made both to suppress electron fake events (which are more forward) and to avoid confusion from initial-state radiation. But the jet E_T cut was made for a fundamentally different reason. At the earliest stage of reconstruction, jets were not found unless they possessed at least 8 GeV of uncorrected E_T. After corrections, the 50%-efficient point rose to 11-12 GeV; even after a cut at 15 GeV was made, there remained concern about modeling the efficiency turn-on. The 8 GeV reconstruction threshold had been invoked to suppress calorimeter noise, but, unfortunately, it was left at the same value for 0.7, 0.5, and 0.3 cones; had cone size been taken into account, the 8 GeV threshold should have been reducible to ∼6 GeV for the 0.5 cone jets used in top mass analysis. (Note that, for the purpose of cuts, in DØ Run 1 top analysis all jets and leptons were considered to have zero invariant mass, so that E_T and p_T were interchangeable.)

Two additional cuts completed the "precut" definition; both came late in the game and neither was straightforward. First was the requirement $E_T^W > 60$ GeV, where E_T^W is the *scalar* sum of the isolated lepton and neutrino transverse momenta. This cut further reduced contamination by electron fakes, which tended to have small electron E_T as well as small missing E_T.

The final unstraightforward "precut" was based on the observation, for lepton + 3-jet events arising primarily from W+jet background, that the η distribution of the W was much flatter for data than for W + 3-jet VECBOS (see Fig. 8(a)). We cut out the region $|\eta^W| > 2$ to improve the agreement, which still wasn't completely acceptable; we speculated that the LO VECBOS was not fully capable of modeling W production in the forward region, where diagrams involving gluons could be important. (When reconstructing the W from the isolated lepton and missing E_T, we used the solution with smaller laboratory W energy; if mismeasurement caused the isolated lepton + missing E_T transverse mass M_W^T to exceed M_W, we identified the W direction with that of the isolated lepton.)

Why is it important to separate top from background without biasing top mass?

At an early stage it was shown (within DØ by D. Chakraborty and others) that a cut on H_T, the scalar sum of the $|p_T|$'s of all jets passing standard cuts as described above, is effective in separating top from the dominant W+jets background. Unfortunately, H_T is strongly correlated with the fit top mass of the event. (This is true regardless of the details of the top mass fitter; analysis described here relied on a "fast" fitter MTL4J, which was *not* used to produce the fit top mass results that DØ published.) This strong correlation has an undesirable practical effect. After a hard cut is placed on H_T, e.g. $H_T > 180$ GeV as in the topological top cross section analysis, the plot of mean fit top mass *vs.* generated top mass flattens considerably, with a slope of order 0.5 rather than 1. To obtain a correct uncertainty in true top mass, one must divide the uncertainty in fit top mass by this slope; a significant loss of sensitivity results. Evidently, to separate top signal from W+jets background, it is preferable to rely on kinematic variables that are not strongly correlated with fit top mass.

In DØ's first publication [3] in which top was observed, the effect of an H_T cut on the fit top mass was so severe that it was decided instead to base the reported mass analysis on a loose-cut sample that unfortunately was dominated by background. This early strategy needed to be improved for further, more quantitative work.

How were unbiased separation variables identified?

Using available GEANTed and RECOed samples of HERWIG 5.7 180 GeV top, W + 4-jet VECBOS interfaced to HERWIG, and a data-derived sample of fake electrons or muons + 4 jets, we

searched for variables that
• discriminated between top and background that had been reconstructed with high top mass ($m_{\rm MTL4J} > 145$ GeV);
• exhibited a (standard, *i.e.* Pearson product moment linear) correlation coefficient $\mathcal{R}$ with $m_{\rm MTL4J}$ that was of order 0.1 or less.

For an initial coarse assessment of the discriminating power of a variable v, we performed an arbitrary crude transformation $f(v)$ such that the distribution of f lay within a finite range and exhibited a single maximum. Then, recording its mean $\langle f_{s,b}\rangle$ and rms $\sigma_{s,b}$ for signal s and background b, we formed a figure of merit

$$F \equiv \frac{|\frac{1}{2}\langle f_s\rangle - \frac{1}{2}\langle f_b\rangle|}{\sqrt{\frac{1}{2}\sigma_s^2 + \frac{1}{2}\sigma_b^2}} \ . \qquad \text{(II.1)}$$

The background samples were cut to satisfy $m_{\rm MTL4J} > 145$ GeV as explained above; the signal events were not tagged by a non-isolated muon.

At this initial stage, scores of variables were considered, many of which were merely small variations upon one another. For example, the variable H_{T2}, defined as H_T less the $|p_T|$ of the leading jet, had about the same F as H_T, but exhibited a slightly reduced (but still unacceptably high) correlation $\mathcal{R}$ with $m_{\rm MTL4J}$. We speculated that, for a given degree of $m_{\rm MTL4J}$ correlation, the leading jet has less discrimination power than do subleading jets because it is not uncommon for a produced W to recoil against a single jet (though with typical E_T only of order $\alpha_S\, M_W$).

What variables were studied most seriously?

The variables considered most seriously were:

$$\begin{aligned}
v_1 &\equiv \text{raw missing } E_T \\
v_2 &\equiv \mathcal{A} \equiv \tfrac{3}{2} \ \times \ \text{least eigenvalue of } \mathcal{P} \\
v_3 &\equiv H'_{T2} \equiv \frac{H_{T2}}{H_\parallel} \\
v_4 &\equiv K'_{T\rm min} \equiv \frac{(\text{min of 6 } \Delta\mathcal{R}_{jj}) \cdot E_T^{\text{lesser } j}}{E_T^W} \\
v_5 &\equiv \eta_{\rm rms} \equiv \sqrt{\langle p_T \text{ weighted } \eta^2\rangle} \ .
\end{aligned} \qquad \text{(II.2)}$$

In the above, "raw" means that the missing E_T was uncorrected for [possible rare] rescaling of an isolated muon momentum in order to reduce the μ-ν transverse mass to the W pole mass (see below). $\mathcal{A}$ is the aplanarity, and $\mathcal{P}$ is the normalized laboratory momentum tensor derived from the momenta of the jets and that of the W. $H_\parallel$ is the scalar $|p_z|$ of the jets, isolated lepton, and neutrino. $\Delta\mathcal{R}_{jj}$ is the distance in η-ϕ space between a pair of jets; of the 6 possible $\Delta\mathcal{R}_{jj}$ belonging to the 4 leading jets, the smallest is chosen. Its factor $E_T^{\text{lesser } j}$ is the smaller of the E_T's (with respect to the beam axis) of the two jets comprising this pair. For v_5, the average extends over the jets and also the W.

One needs to solve for the W momentum in order to calculate v_2, v_3, v_4, and v_5. Normally, the W was reconstructed from the isolated lepton momentum and the missing E_T; two solutions were found, and the one with smaller laboratory W energy was retained. Occasionally, no solutions were found because the lepton+neutrino transverse mass M_T^W exceeded the W pole mass M_W. In this case, provided that the isolated lepton was not a muon with $p_T > 85$ GeV, the neutrino rapidity was set equal to that of the isolated lepton and M_T^W was allowed to continue to exceed M_W. This choice took account of the fact that the W lineshape does have a Breit-Wigner tail. On the other hand, if the isolated lepton was a muon with $p_T > 85$ GeV, it did not lie far from its maximum detectable momentum (where $\sigma(1/p) = 1/p$), so the large M_W^T was likely due to gross mismeasurement of the muon momentum. In that case we reduced p_μ, keeping its direction fixed, and concomitantly recalculating the missing E_T, until $M_W^T = M_W$.

To do something more rational in Run 2, one realizes in hindsight that DØ needs WFITTER, written in the spirit of ZFITTER used in Run 1. WFITTER would accept as input the isolated lepton momentum and missing E_T and their errors (preferably expressed in terms of $1/p$ for muons). Taking into account the known W lineshape in addition to these errors, WFITTER would return the most likely W four-momentum, its errors, and, if $M_W^T > M_W$, a measure of the probability

that in fact the input information is consistent with W decay.

Do these variables separate top from background?

We discuss the ability of these variables to separate top from background, taking the least complex variables first. In all figures, "top" refers to HERWIG 5.7 with a top mass of 180 GeV; "bkgnd" or "W+jets" refer to a combination of a major background consisting of W + 4-jet VECBOS showered by HERWIG, and a minor background consisting of data-derived 5-jet events in which one jet is electron-like. The ratio of minor to (major+minor) backgrounds was fixed at 0.18. After this cut, all background events that are plotted were required to have $m_{\mathrm{MTL4J}} > 145$ GeV; taking into account the MTL4J calibration, this corresponds to a ∼150 GeV cut. In all histograms, top and background are normalized to the same area; horizontal axes are labeled by bin *center*, with labels centered on the appropriate bin; and underflows and overflows are included in the extreme bins.

Event aplanarity variable

Figure 1 exhibits the top and background distributions in a function of $\mathcal{A} \equiv v_2$, the aplanarity.

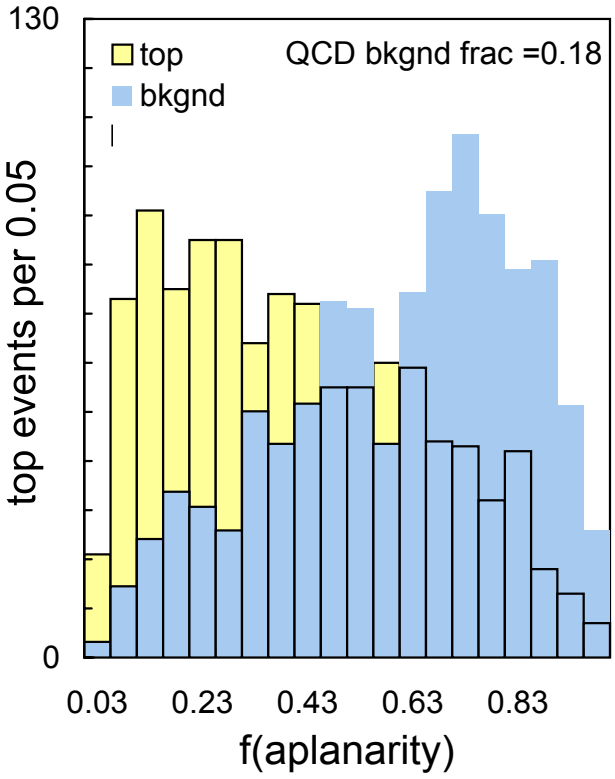

FIG. 1. Top and background distributions in $f(\mathcal{A})$, where $\mathcal{A} \equiv v_2$ is the aplanarity and the functional form plotted is $f(\mathcal{A}) = \exp(-11\mathcal{A})$. (Horizontal bin-center labels are rounded to two significant figures, e.g. 0.03 should be read as 0.025.)

Note that the objects contributing to the momentum tensor used to calculate $\mathcal{A}$ include the W as well as the jets, and that $\mathcal{A}$ is computed in the *lab*, not in the C.M. While it made good sense (and it slightly improved the discrimination) to include the W in this measure of event shape, it made less sense to use the lab frame for a shape variable that historically was invented for use in the C.M. Once the W was reconstructed, it would have been easy to boost the event along the z axis so that the net z momentum vanishes. If it had been calculated in this boosted frame, $\mathcal{A}$ would have been unaffected by irrelevant fluctuations arising from random longitudinal boosts of the quark-antiquark system. [A counterargument claimed that $t\bar{t}$ events, because of their larger $\sqrt{\hat{s}}$, are produced with smaller longitudinal boost than are background events, giving $\mathcal{A}$ more discrimination power when

it is computed in the lab. But this argument didn't take into account the availability of specific measures of event centrality, such as H'_{T2} and $\eta_{\rm rms}$ discussed below, that are particularly effective in rejecting highly boosted events.]

Jet centrality/transverse energy variable

Figure 2 exhibits the top and background distributions in a function of $H'_{T2} \equiv v_3$, where $H'_{T2} \equiv H_{T2}/H_{\|}$, $H_{T2} \equiv H_T - |p_T^{\text{leading jet}}|$, and

$$H_T \equiv \sum_{\text{jets}} |p_T|$$
$$H_{\|} \equiv \sum_{\text{jets},\ell,\nu} |p_z| \ . \qquad \text{(II.3)}$$

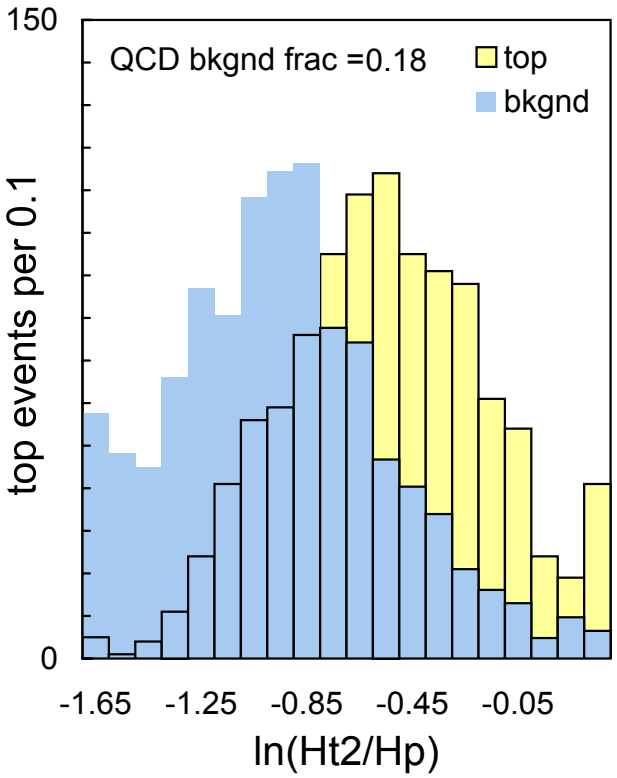

FIG. 2. Top and background distributions in $\ln(H'_{T2}) \equiv \ln v_3$ *(see text).*

We have already discussed why we prefer H_{T2} to H_T as a jet E_T variable for defining a top mass analysis sample. But why do we choose to divide H_{T2} by $H_{\|}$? The issue here is correlation with $m_{\rm MTL4J}$. Aplanarity, an event shape variable, is largely uncorrelated with $m_{\rm MTL4J}$, while H_{T2}, a jet transverse energy variable, is strongly ($\sim \frac{1}{2}$) so correlated. This is expected: higher-E_T jets yield higher top masses. If we require an H_T-like variable that is largely uncorrelated with $m_{\rm MTL4J}$, we must modify H_{T2}. Here we choose to divide it by a second variable, $H_{\|}$, that also scales with object energies and thus is also strongly correlated with $m_{\rm MTL4J}$. Fortunately, $H_{\|}$ typically is larger for high-$m_{\rm MTL4J}$ background than for signal, so its effect in the denominator enhances rather than degrades the discriminating power of the quotient. This point is illustrated by plotting in Fig. 3 the top and background distributions for H_{T2} itself.

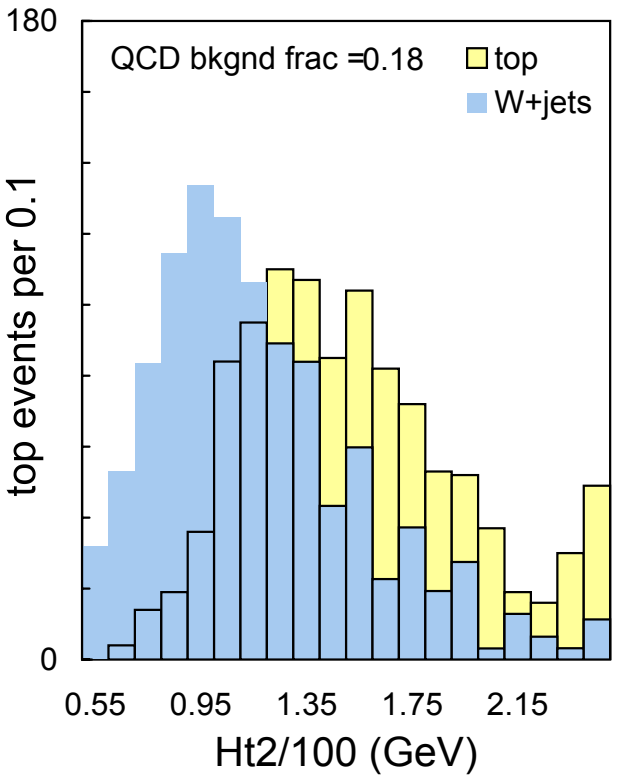

FIG. 3. Top and background distributions in H_{T2} (GeV)/100.

Qualitatively comparing Figs. 2 and 3, it does not appear that H_{T2} has a discriminating power (against background with $m_{\rm MTL4J} > 145$ GeV) that is superior to that of H'_{T2}. With this definition, H'_{T2} primarily measures jet *centrality* rather than transverse energy.

If dividing a variable similar to H_T by $H_{\|}$ is a good idea, why wasn't this done in the

cross section analysis? One reason is that the background-discrimination problems for the mass and cross section analyses aren't strictly comparable; for mass analysis one strives to reject background that is reconstructed with high apparent top mass, while for cross section analysis one is most interested in rejecting all background. It should be noted that the cross section analysis didn't make use of a variable designed explicitly to select central events, although it did make cuts on the maximum $|\eta|$ of objects.

Event centrality variable

Figure 4 exhibits the top and background distributions in $\eta_{\rm rms} \equiv v_5$, where

$$\eta_{\rm rms} \equiv \sqrt{\frac{\sum_{W+4\rm jets} |p_T|\eta^2}{\sum_{W+4\rm jets} |p_T|}} \ , \qquad \text{(II.4)}$$

and the sum is taken over the four leading jets plus the W.

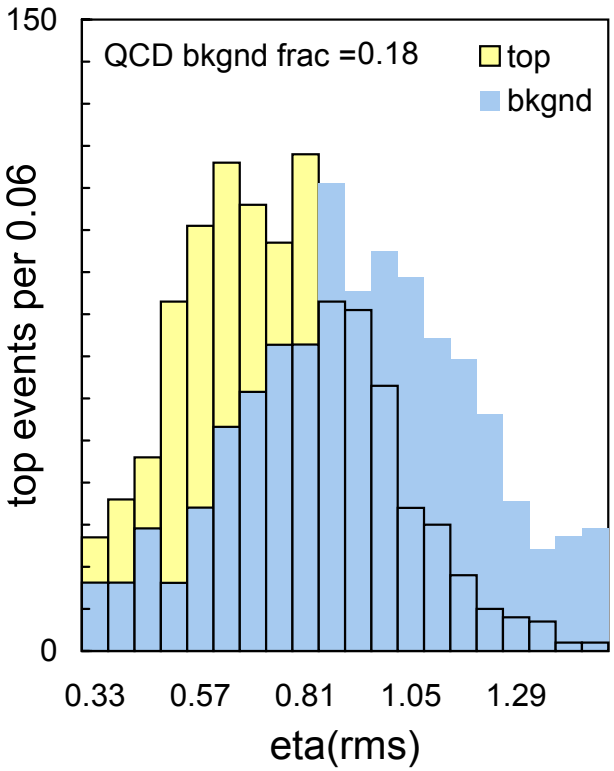

FIG. 4. Top and background distributions in $\eta_{\rm rms} \equiv v_5$ (see text).

After experimentation, it was found to be slightly better to include the W in the sum; to exclude extra jets beyond the four leading ones; to compute the root-mean-square rather than the mean of $|\eta|$; and to weight η^2 by $|p_T|$ to the first rather than to another power. Like $\mathcal{A}$ and H'_{T2}, $\eta_{\rm rms}$ is only weakly correlated with $m_{\rm MTL4J}$.

Jet separation variable

Figure 5 exhibits the top and background distributions in a function of $K'_{T\rm min} \equiv v_4$, where $K'_{T\rm min} = K_{T\rm min}/E_T^W$, and E_T^W, the sum of the $|p_T|$'s of the isolated lepton and the neutrino, is the same variable that was required to exceed 60 GeV in the precut sample definition. $K'_{T\rm min}$ is equal to the smallest of the $\Delta\mathcal{R}_{jj}$'s that can be formed from the 6 pairs jj of the 4 leading jets, multiplied by $E_T^{\rm lesser\ j}$, the smaller of the jet E_T's of the 2 jets belonging to this smallest-$\Delta\mathcal{R}_{jj}$ pair.

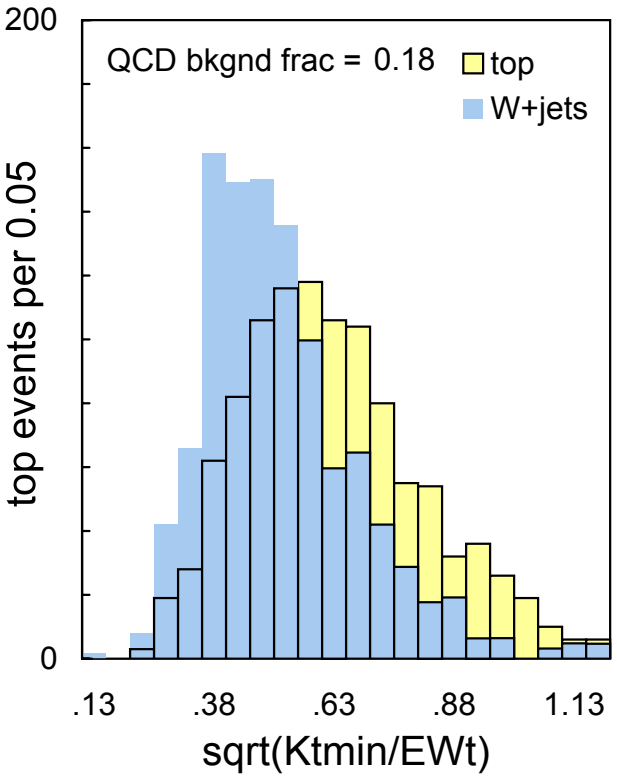

FIG. 5. Top and background distributions in $\sqrt{K'_{T\rm min}} \equiv \sqrt{v_4}$ (see text).

$K_{T\rm min}$ probes the origins of the four leading jets in the final state. For top events these jets arise from decay of heavy objects and are usually distinct, with ΔR_{jj} not small. On the other hand,

for W+jets and QCD background, the extra jets arise from radiation, which can cause pairs of jets to cluster. Multiplying ΔR_{jj} by $E_T^{\text{lesser } j}$ allows $K_{T\min}$ to approximate a momentum transverse to the radiation axis, which is a key variable controlling the radiation probability.

After experimentation, it was found to be simpler, and as effective, to consider only the two jets closest in ΔR, and to raise ΔR_{jj} and $E_T^{\text{lesser } j}$ to the same power.

In analogy to the evolution of H'_{T2} from H_{T2}, the description of $K_{T\min}$ would end here, except for its significant ($\sim \frac{1}{4}$) correlation with m_{MTL4J}, not unexpected in view of the transverse energy factor that it includes. Again we sought to mitigate this correlation. If again one were to divide by $H_{\parallel}$, one would introduce too much mutual correlation with H'_{T2}. Instead we divided $K_{T\min}$ by E_T^W, which is similarly correlated with m_{MTL4J}; this defined $K'_{T\min}$.

We were less pleased with $K'_{T\min}$ than we were with H'_{T2}. Our main reservation concerning H'_{T2} was that dividing H_{T2} by $H_{\parallel}$ transformed a measure of jet transverse energy to a measure of jet centrality (thereby introducing unwanted correlation with other variables like η_{rms} that measure event centrality). Otherwise we were happy with H'_{T2}'s performance and cleanliness of definition.

Unfortunately, $K'_{T\min}$ is not as cleanly constructed – we divide a jet-related quantity by a lepton-related quantity, which seems to make sense only as a means of reducing the top mass correlation. Furthermore, doctoring $K_{T\min}$ in this way slightly reduced its rejection power. This is seen by comparing Fig. 5 with Fig. 6, which exhibits the top and background distributions in a function of the undoctored $K_{T\min}$.

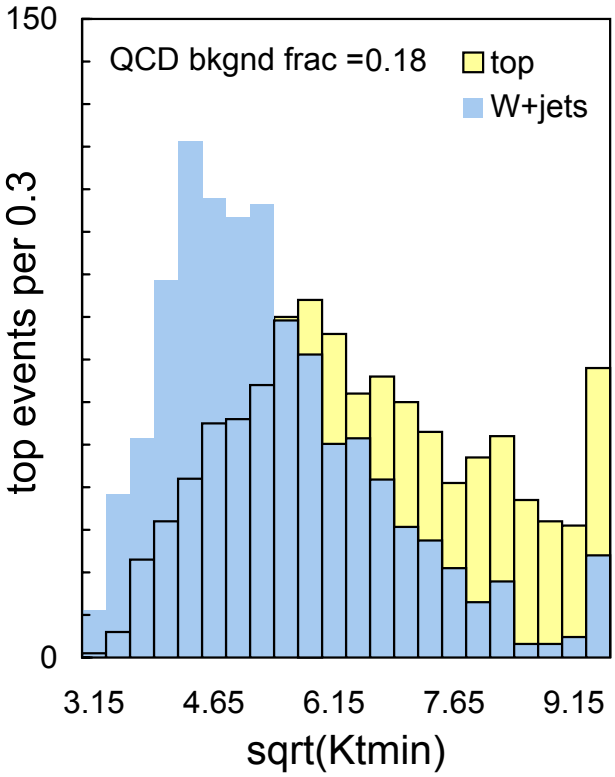

FIG. 6. Top and background distributions in $\sqrt{K_{T\min}}$ *(see text).*

Notwithstanding these complaints, we lacked a better idea for reducing the top mass correlation of $K_{T\min}$, so we proceeded with use of $K'_{T\min}$.

In Run 2-inspired hindsight, are there further thoughts on $K'_{T\min}$? One modest point is that, in a small-$\Delta\eta$ small-$\Delta\phi$ approximation, the familiar variable p_T^{rel}, representing the momentum of either (massless) jet 1 or jet 2 perpendicular to the sum of their momenta $\vec{p}_1 + \vec{p}_2$, is given by

$$p_T^{\text{rel}} \approx \Delta\mathcal{R}_{12} \frac{p_{T1} p_{T2}}{p_{T1} + p_{T2}} \,. \qquad \text{(II.5)}$$

Setting $K_{T\min} \equiv p_T^{\text{rel}}$ as defined by this particular equation may not have been tried in Run 1 analysis; it might be a slightly better choice than simply multiplying $\Delta\mathcal{R}_{12}$ by the smaller of p_{T1} and p_{T2}. If so, one should check whether the jet pair would better be chosen on the basis of minimum p_T^{rel} rather than minimum $\Delta\mathcal{R}$.

Missing energy variable

Figure 7 exhibits the top and background distributions in a function of missing $E_T \equiv v_1$. This

variable is computed using the calorimeter and the muon system (not the calorimeter alone). As mentioned earlier, here we call it "raw" because it is uncorrected for the (rare, previously described) rescaling of p_μ and p_T^ν that was carried out when $M_T^W > M_W$ and $p_T^\mu > 85$ GeV. The function $f(v_1)$ for which these distributions are plotted is given in the caption; it is different for e+jets and μ+jets events because both the v_1 threshold and the v_1 resolution are different for these two classes.

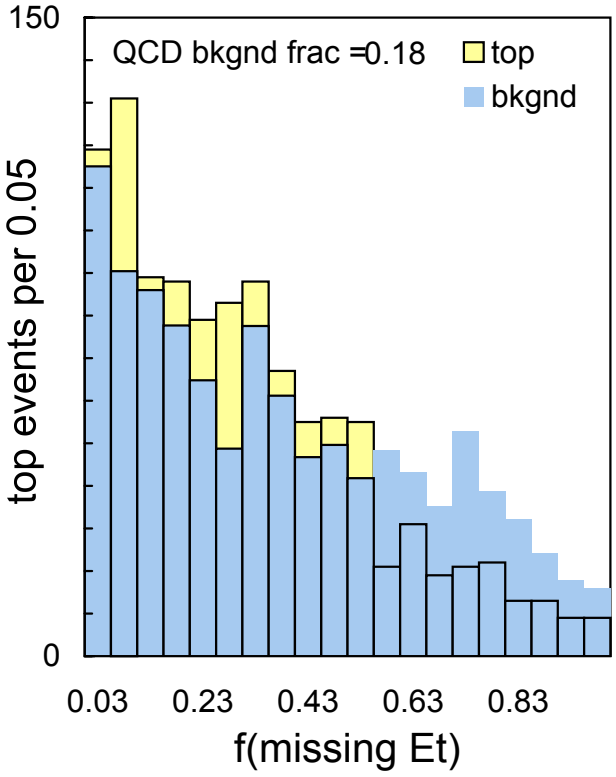

FIG. 7. Top and background distributions in $f(v_1)$, where $v_1 \equiv$ missing E_T (in GeV), and, for ℓ+jets events,

$$f(v_1) \equiv \exp\left(-\max\left(0, ((\sqrt{v_1} - c_1)/c_2)\right)\right)$$
$$c_1 = \sqrt{25},\ c_2 = 1.5\ (\ell = e)$$
$$c_1 = \sqrt{20},\ c_2 = 3.0\ (\ell = \mu)\ . \tag{II.6}$$

Figure 7 is qualitatively different from the first 6 figures, showing much less discrimination between top signal and background. Why did we consider v_1 at all?

To explain, we refer to some history. At an earlier stage of analysis progress, when fake non-isolated-lepton "QCD" background was a larger issue, we sought to distinguish among all 3 of the event classes – top, W+jets, and QCD. To each event, we aimed to assign probabilities $P_{\rm top}$, $P_{W+\rm jets}$, and $P_{\rm QCD}$, with their sum equal to unity. The 3 probabilities for each event could displayed as a point on the surface of an equilateral triangle, with each vertex representing unit probability for a particular class. Events accepted for top mass analysis were to be classified by drawing an optimized contour on the triangle that included its top vertex.

Correspondingly, we sought to develop variables that were effective in distinguishing each class from the others; in particular, QCD events exhibited low missing E_T relative to W+jets and top events, which include genuine neutrinos. It was in this context that the variable v_1 was defined and used.

At a later stage of top analysis, after the QCD background was further suppressed by better isolated lepton identification, we reexamined the effectiveness of this "triangle" scheme in competition with a simpler scheme in which both backgrounds were lumped together. Using top efficiency *vs.* purity as a criterion, we found the two methods to be essentially the same. From that point onward, we focused on separating top from a single background sample consisting of 82% W+jets and 18% QCD. In that context, as emphasized by Fig. 7, v_1 provided only very limited discrimination.

As will be described in Section IV, we developed a top likelihood using a quantitative method for determining whether and with what optimum weight each potential variable should be used. Despite the very modest discrimination power of v_1, this method determined that it is better to include v_1 in the likelihood than to leave it out.

Was DØ able to model these variables?

In the lepton+jets top mass PRL [4], we addressed this question using W+3 jet control samples as illustrated in Fig. 8 (c–f). The PRL text stated "As shown for the background dominated W+3 jet sample in Fig. 8 (c–f), x_1–x_4 [v_1–v_4 in the notation of the present note] are reasonably well modeled by MC; this is true also for the W+2 jet and top mass samples (not shown)."

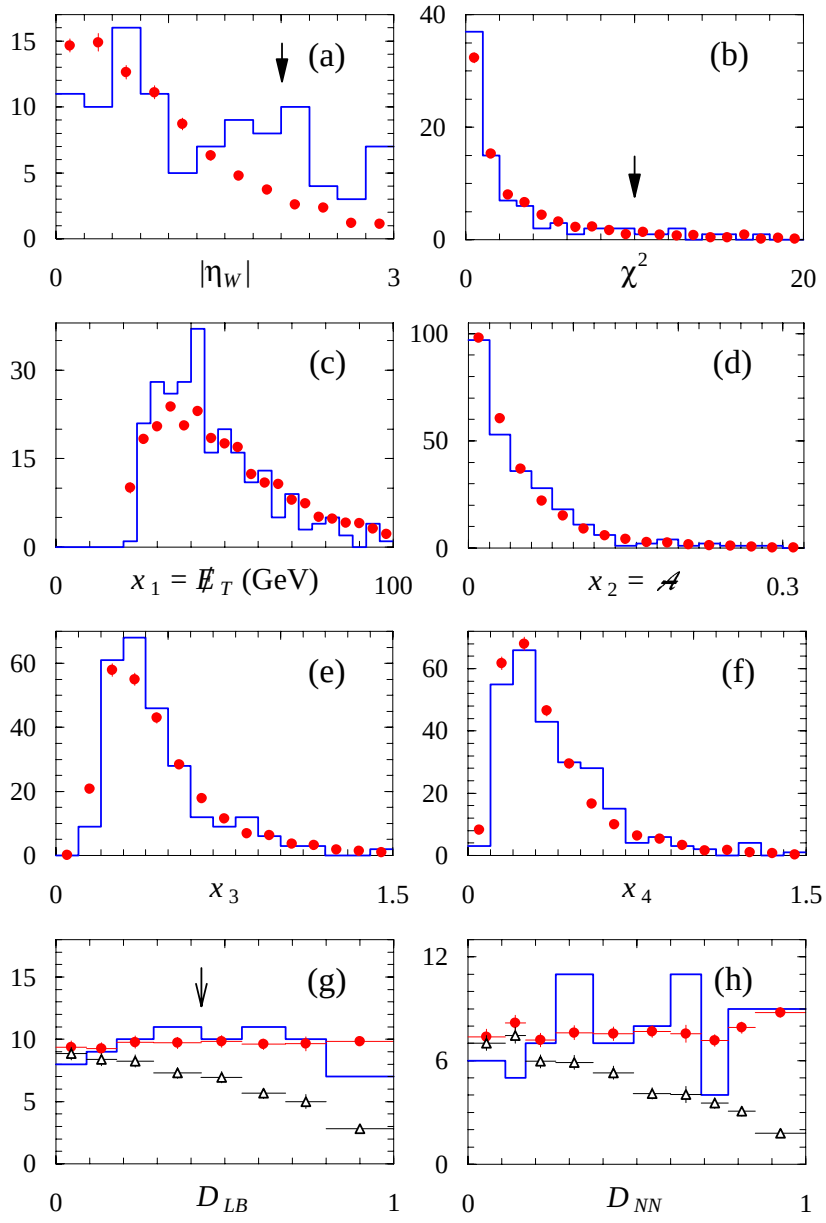

FIG. 8. (This reproduces the PRL caption). "Events per bin vs. event selection variables defined in the text, plotted for (a–b, g–h) top quark mass analysis samples, and (c–f) W+3 jet control samples. Histograms are data, filled circles are expected top + background mixture, and open triangles are expected background only. Solid arrows in (a–b) show cuts applied to all events; the open arrow in (g) illustrates the LB cut. The nonuniform bin widths in (g–h) are chosen to yield uniform bin populations."

The basic idea in Fig. 8(c–f) was to use W + 3-jet events, untagged by nonisolated muons and largely background, as a test of DØ's ability to model variables v_1–v_4 in its background sample. As is seen from the plots, the agreement wasn't perfect, but statistics were limited. After the paper was submitted, lively correspondence occurred between DØ and one of the referees. The following is a fragment from that correspondence:

"*Referee B continues, 4. At the bottom of page 7, the x_i's are introduced. The last two of these, x_3 and x_4, have potential sensitivity to the modeling of gluon radiation in the Monte Carlo. Are there tests in the data that confirm the Monte Carlo and give a measure of the systematic uncertainty?*

Reply: In the background dominated W + 3 jet sample shown in Figs. 1(e) and (f), Monte Carlo matches data with chisq_y probabilities of 14% and 2%, respectively. The same distributions are in general agreement for the W + 2 jet sample, and in detailed agreement for the W + 4 jet sample upon which this analysis is based. The systematic error includes +-100% of the effect of changing the VECBOS QCD scale from jet ⟨p_T⟩ˆ2 to M_Wˆ2, and +-100% of the effect of substituting the ISAJET Monte Carlo generator for HERWIG in fragmenting the VECBOS partons.

Detail: Figures 1(e) and (f) present one set of tests that quantify the reliability of the VECBOS MC simulation of x_3 and x_4. Here the sample is W + 3 jets, which are dominated by background and which have much higher statistics than the W + 4 jet data upon which the top mass measurement is based. For this sample and these tests, the chisq_y probabilities are 14% and 2%, respectively

As mentioned on Page 7 of the manuscript, the same set of tests was repeated for the W + 2 jet and W + 4 jet final states. For the W + 2 jet tests, the data and simulation are in general agreement; for the distributions that are analogous to Figs. 1 (e) and (f), the ratio of means for simulation and data are 0.94 and 0.92, respectively, with an error of about 2.5%. For the W + 4 jet final states, which are the direct basis for this analysis, the data and Monte Carlo are in excellent agreement.

As stated above and on Page 10, we included in the systematic error estimate +-100% of the effect of changing the VECBOS QCD scale from jet ⟨p_T⟩ˆ2 to M_Wˆ2, and +-100% of the effect of substituting the ISAJET Monte Carlo generator for HERWIG in fragmenting the VECBOS partons. These are substantial excursions of the background model, based on variations that we observe in the level of agreement between W + jets data and the simulated background for many different distributions. Allowance for these systematic errors in the cross-checks performed with the W + 3 jets events would further enhance the degree to which the agreement between data and simulation in those tests is considered to be satisfactory.

*Referee B continues,This raises the general question of determining the systematic uncertainty in the multivariate analysis. For example, the data and Monte Carlo distributions of the missing ET (fig. 1c) are qualitatively similar, but quantitatively they are not the same (large chi**2 without the first point, very large with it). How does this propagate into a systematic uncertainty?*

Reply: If there were a problem in modeling the background to which the calculation of D is especially sensitive, this would be manifested most strongly as an inconsistency between data and simulation in the distribution of D itself. Figs. 1(g) and (h) confirm that the distribution of D_LB and D_NN in the data is consistent with the simulation; the chisq_y probabilities are 96% and 64%, respectively.

Detail: For Fig. 1c, the variable chisq_y (defined above) is 26.1 (33.7) for 18 (19) degrees of freedom, if the first populated bin is not (is) included, yielding a chi-square probability of 10% (2%). The W + 3 jet control samples used to construct this cross-check play no direct role in the top mass analysis.

For the analogous plot using the W + 2 jet top samples, general agreement is obtained; the means of the simulated and data distributions are mutually consistent, and the RMS's are the same within 6% (1.2 standard deviations). For

the analogous plot using the W + 4 jet top mass samples, which do play a direct role in the top mass analysis, the agreement is superb... "

Note that "*chisq_y*" is defined by

"*Here N is the no. of data counts and y is the simulation (normalized to the data). Then chisq_y is sum (N_i-y_i)^2/(y_i+dy_i^2) (dy_i is the error on y_i)* ."

To summarize this correspondence, in Run 1 DØ argued that the distributions of v_1–v_4 were adequately modeled; ultimately, PRL accepted that argument. At that time, would DØ have been able to argue that the *correlations* among these variables were adequately modeled? In the author's opinion, more QCD statistics, extension of VECBOS to NLO (so that *e.g.* η_W is better predicted), and serious tuning of Monte Carlo to detector performance would have been required in order to do a convincing job of that.

III. NEURAL NETWORK DISCRIMINANT

As mentioned in the Introduction, in two different analyses DØ used two different discriminants to separate top from background: a likelihood discriminant $\mathcal{D}_{\rm LB}$ and a neural network (NN) discriminant $\mathcal{D}_{\rm NN}$. $\mathcal{D}_{\rm NN}$ was computed primarily by P. Bhat and H. Prosper; while the details of their work lie beyond the scope of the present discussion, $\mathcal{D}_{\rm NN}$ retains a strong connection to the present note, since variables v_1-v_4 were the NN inputs.

Here we mention the following points concerning $\mathcal{D}_{\rm NN}$:

• The classic difference between likelihood- and NN-based discriminants is that the latter are sensitive to (and can gain discrimination from) differences between signal and background in the *correlations* among input variables. As mentioned above, in Run 1 DØ argued that it was able to model the background and signal distributions in v_1-v_4. At that time, DØ didn't argue correspondingly that it was able to model the correlations among those variables.

• A "kitchen-sink" NN approach to selecting a lepton+jets sample for top mass analysis would use as NN input a collection of familiar variables, such as H_T and $H_\parallel$; individual jet, lepton, and neutrino p_T's; shape variables like aplanarity and centrality; and a measure of the number of jets. At an early stage, variations on this approach were attempted; unfortunately, they yielded discriminants that weren't useful for top mass analysis. The accepted events were skewed strongly toward high apparent top mass. The accepted background sample's reconstructed top masses rose to values near that of the signal; samples with disparate generated top masses exhibited reconstructed top masses that were more tightly bunched. These effects compromised both statistical and systematic errors. Only after sets of nearly top-mass-uncorrelated variables such as v_1-v_4 were used as input did the NN-selected samples become useful.

• Taking an idealized view of Run 2, assuming that there will be no strong limitations on Monte Carlo accuracy or background statistics, one would expect from the "kitchen-sink" approach greater statistical power in separating top from background than could be obtained with the limited set of variables described in this note. For example, consider our variable $v_3 = H_{T2}/H_\parallel$. Clearly more information is contained in the separate values of H_{T2} and $H_\parallel$ than in their quotient alone. Can DØ identify a strategy that would unlock any additional discrimination without biasing the apparent top mass of the selected sample? Two approaches come to mind:

◦ Identify or invent a type of NN that can be trained to produce an output $\mathcal{D}$ which discriminates maximally between two sample types, subject to the constraint that $\mathcal{D}$ is uncorrelated with a particular additional variable.

◦ Train an NN on top background and signal samples, reweighting the background to match exactly the apparent top mass distribution of the signal. (This would require an NN that is capable of being trained on weighted events.) We speculate that, if trained in this way, the NN would yield an output that is uncorrelated with the apparent top mass.

IV. LIKELIHOOD DISCRIMINANT

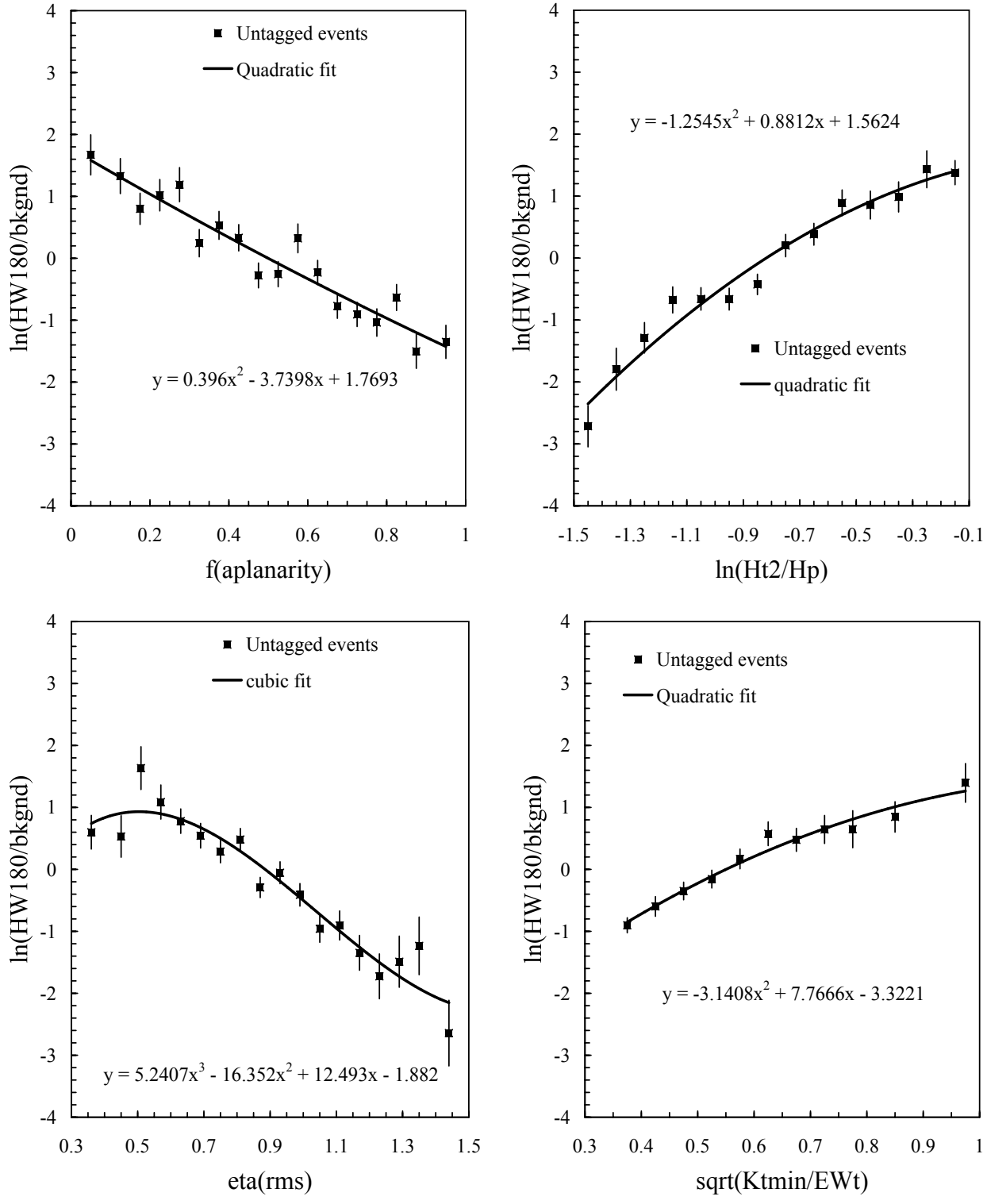

FIG. 9. Natural logarithm of ratio of top to background for the distributions plotted in Fig. 1 ($f(v_2)$, top left), Fig. 2 ($\ln v_3$, top right), Fig. 4 (v_5, bottom left), and Fig. 5 ($\sqrt{v_4}$, bottom right).

Fitting the top/background ratios

With variables $\{v_i\}$ defined, the first step in forming the likelihood discriminant $\mathcal{D}_{\rm LB}$ was to plot bin-by-bin the ratio

$$L_i^{tb} \equiv \ln\,(\text{top/background})$$

for the distributions in $x \equiv f(v_i)$ shown in Figs. 1, 2, 4, 5, and 7, and to fit the result to a quadratic or cubic $y(x)$. Figures 9 and 10 exhibit these ratios and fits.

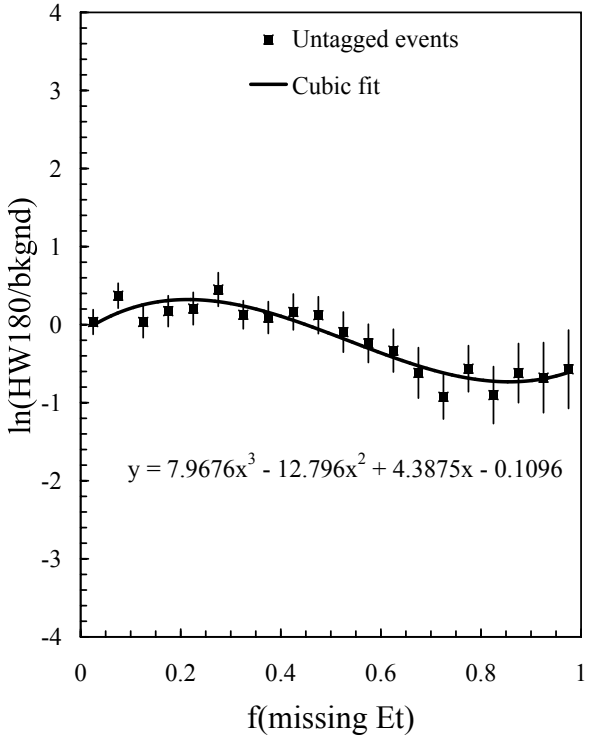

FIG. 10. Natural logarithm of ratio of top to background for the distributions plotted in Fig. 7 ($f(v_1)$).

Though the quadratic and cubic fit results displayed in Figs. 9 and 10 seem straightforward enough, pitfalls were avoided by respecting the following points:

• When fitting the ratio of two distributions, the distributions should be divided first and fit afterward, rather than the reverse. Otherwise small variations in the initial fits near the distributions' tails can cause large variations in the ratios.

• The logarithm of the ratio should be fit, not the ratio itself. This simplifies the fit function and symmetrizes the errors on the points.

• To prevent "error of the error" biases, bins should be adequately populated. (In the ratio involving $\sqrt{v_4}$, two interior bins were combined.)

• To avoid bias, no under/overflow points should be ignored. Instead, data near the tail of a distribution should be consolidated in a single bin with adequate statistics and plotted at an appropriate abscissa.

• The indicated quadratic or cubic fits should be used only within the range of the points that were fit. Those ranges were

$$\begin{aligned} 0 \le f(v_1) &\le 1 \\ 0 \le f(v_2) &\le 1 \\ -1.5 \le \ln v_3 &\le -0.1 \\ 0.35 \le \sqrt{v_4} &\le 1 \\ 0.36 \le v_5 &\le 1.50\ . \end{aligned} \qquad \text{(IV.1)}$$

If the abscissa lies beyond that range, the quadratic or cubic should be evaluated at the range limit.

Correlations between top mass and discriminant variables

Up to now, this note has remained vague about the central feature of variables v_1-v_5 - that they are only weakly correlated with $m_{\rm MTL4J}$. This is deliberate: up to now L_i^{tb} had not been introduced, while the relevant correlation is the one between $m_{\rm MTL4J}$ and L_i^{tb}, not between $m_{\rm MTL4J}$ and v_i itself. This is because L_i^{tb} rather than v_i measures the discrimination between top signal and background, and it is this discrimination we wish to be uncorrelated with $m_{\rm MTL4J}$. As usual, the correlation C_{mL_i} between $m \equiv m_{\rm MTL4J}$ and $L_i \equiv L_i^{tb}$ is expressed by the usual Pearson product moment linear correlation coefficient

$$\mathcal{C}_{mL_i} \equiv \frac{\langle (m-\bar{m})(L_i-\bar{L}_i)\rangle}{\sigma_m \sigma_{L_i}}\ , \qquad \text{(IV.2)}$$

where *e.g.* σ_{L_i} is the root variance of L_i.

Measured on W+jets-only background (with $m_{\rm MTL4J} > 145$ GeV, as usual), the values of

$\mathcal{C}_{mL_i^{tb}} \equiv \mathcal{C}_{mi}$ were

$$\begin{aligned} 0.019 &= \mathcal{C}_{m1}\ (\text{missing } E_T) \\ -0.082 &= \mathcal{C}_{m2}\ (\mathcal{A}) \\ 0.017 &= \mathcal{C}_{m3}\ (H'_{T2}) \\ 0.055 &= \mathcal{C}_{m4}\ (K'_{T\min}) \\ 0.040 &= \mathcal{C}_{m5}\ (y_{\rm rms}) \end{aligned} \qquad \begin{aligned} 0.432 &= \mathcal{C}_{m6}\ (H_{T2}) \\ 0.239 &= \mathcal{C}_{m7}\ (K_{T\min})\ . \end{aligned} \tag{IV.3}$$

At the bottom are included the correlations with $m_{\rm MTL4J}$ of two variables that were not used in the discriminant because these correlations were too large.

Correlations among discriminant variables

Tabulated are the correlations $\mathcal{C}_{ij}$ measured between pairs of discriminant variables L_i^{tb} and L_j^{tb}:

$$\mathcal{C}_{ij} = \begin{pmatrix} & {\scriptstyle 2} & {\scriptstyle 3} & {\scriptstyle 4} & {\scriptstyle 5} \\ {\scriptstyle 1} & -0.011 & 0.020 & -0.093 & 0.056 \\ {\scriptstyle 2} & & 0.399 & 0.247 & 0.270 \\ {\scriptstyle 3} & & & 0.132 & 0.771 \\ {\scriptstyle 4} & & & & -0.140 \end{pmatrix} \tag{IV.4}$$

The indicated values are the means of the correlations measured on top signal and on W+jets-only background, which were similar. (Obviously the full correlation matrix $\mathcal{C}$ is symmetric with unit diagonal elements.)

Some of the correlations among these variables are substantial, particularly between H'_{T2} and $y_{\rm rms}$ (0.771) or $\mathcal{A}$ (0.399). These correlations must be taken into account when constructing an optimized discriminant from the set of L_i^{tb}'s.

Log likelihood significance and normalization

Two last components are needed to form the likelihood discriminant $\mathcal{D}_{\rm LB}$ The first is the signal-to-noise ("S/N") s_i, defined for each variable i in gaussian approximation by

$$s_i \equiv \frac{\frac{1}{2}\langle L_i^{tb}\rangle_{\rm top} - \frac{1}{2}\langle L_i^{tb}\rangle_{\rm bkgnd}}{\sqrt{\frac{1}{2}\sigma^2(L_i^{tb})_{\rm top} + \frac{1}{2}\sigma^2(L_i^{tb})_{\rm bkgnd}}}\ , \tag{IV.5}$$

where σ^2 is the variance. Thus s_i is the number of standard deviations by which $\langle L_i^{tb}\rangle$ for an "average" event (top or W+jets-only background) is nonzero. It measures the significance of the log likelihood that is associated with each variable. The values of the elements of the S/N vector $\vec{s}$ were

$$\begin{aligned} 0.110 &= s_1\ (\text{missing } E_T) \\ 0.421 &= s_2\ (\mathcal{A}) \\ 0.492 &= s_3\ (H'_{T2}) \\ 0.369 &= s_4\ (K'_{T\min}) \\ 0.366 &= s_5\ (y_{\rm rms}) \end{aligned} \qquad \begin{aligned} 0.465 &= s_6\ (H_{T2}) \\ 0.400 &= s_7\ (K_{T\min})\ . \end{aligned} \tag{IV.6}$$

These values quantify the assertions, made previously in this document, that H'_{T2} is more efficient than H_{T2}, but $K'_{T\min}$ is less efficient than $K_{T\min}$; and also that missing E_T is much less efficient (for separation of top from W+jets) than the other variables. Note also that H'_{T2} is more efficient than $y_{\rm rms}$, with which it is highly correlated.

The second component, needed for normalization purposes, is merely the numerator of $\vec{s}$:

$$\ell_i \equiv \tfrac{1}{2}\langle L_i^{tb}\rangle_{\rm top} - \tfrac{1}{2}\langle L_i^{tb}\rangle_{\rm bkgnd}\ . \tag{IV.7}$$

For top and W+jets-only background, the values of the elements of the the average log-likelihood vector $\vec{\ell}$ were

$$\begin{aligned} 0.035 &= \ell_1\ (\text{missing } E_T) \\ 0.329 &= \ell_2\ (\mathcal{A}) \\ 0.504 &= \ell_3\ (H'_{T2}) \\ 0.225 &= \ell_4\ (K'_{T\min}) \\ 0.288 &= \ell_5\ (y_{\rm rms}) \end{aligned} \qquad \begin{aligned} 0.449 &= \ell_6\ (H_{T2}) \\ 0.296 &= \ell_7\ (K_{T\min})\ . \end{aligned} \tag{IV.8}$$

Total S/N for uncorrelated variables

Consider, for example, variables v_1-v_4. Hypothetically, if the $\{L_i^{tb}, i = 1, 4\}$ were all mutually uncorrelated, for a given event the combined log likelihood of top *vs.* background would be given by their sum:

$$L^{tb} = \sum_{i=1}^{4} L_i^{tb}\big(f_i(v_i)\big) \ , \qquad \text{(IV.9)}$$

where f_i is the function that was applied to v_i before the ratio of top to background distributions was taken, *e.g.* $f_2(v_2) = \exp(-11v_2)$ (see Fig. 1). (This is a special case of forming the weighted sum

$$L^{tb} = \sum_{i=1}^{4} e_i \, L_i^{tb}\big(f_i(v_i)\big) \ , \qquad \text{(IV.10)}$$

with all weights e_i equal to unity.)

In this hypothetical mutually uncorrelated case, the S/N $s_{\rm tot}$ of the resulting L^{tb} would be

$$s_{\rm tot} = \sqrt{\sum_{i=1}^{4} s_i^2} = 0.753 \ . \qquad \text{(IV.11)}$$

That is, on average, either a top or a background event would have a discriminant that is $\frac{3}{4}\sigma$ different from zero, and $\frac{3}{2}\sigma$ different from the complementary assignment.

Cross-check on gaussian approximation

As soon as one makes use of its correlation matrix $\mathcal{C}$, implicitly one is assuming that L_i^{tb} has a gaussian distribution. A cross-check of this assumption may be made. If the L_i^{tb} are uncorrelated log likelihoods, they should be added with unit weight. On the other hand, if the L_i^{tb} are uncorrelated gaussian-distributed variables measuring the same property (top *vs.* background), it is easy to show that they should be combined with relative weights proportional to the quotient s_i^2/ℓ_i, as defined above. For these two statements to be consistent, s_i^2/ℓ_i should be the same for all i. Plugging in the numbers tabulated above, one obtains (for top and W+jets-only background)

$$\begin{aligned}
0.344 &= s_1^2/\ell_1 \ (\text{missing } E_T) \\
0.540 &= s_2^2/\ell_2 \ (\mathcal{A}) \\
0.481 &= s_3^2/\ell_3 \ (H'_{T2}) \\
0.605 &= s_4^2/\ell_4 \ (K'_{T\rm min}) \\
0.466 &= s_5^2/\ell_5 \ (y_{\rm rms}) \\
\\
0.482 &= s_6^2/\ell_6 \ (H_{T2}) \\
0.541 &= s_7^2/\ell_7 \ (K_{T\rm min}) \ .
\end{aligned} \qquad \text{(IV.12)}$$

The mutual agreement is coarse; in particular, missing E_T departs by $\sim 30\%$ from the norm. Note that the function $f(v)$ against which the ratio of top to background distributions is plotted doesn't affect this check. Fortunately, missing E_T will be seen to play so small a role that the departure of L_1^{tb} from gaussian behavior should not pose a serious concern.

In the previous section, when we were doing the simplest thing, *i.e.* adding uncorrelated likelihoods, we used unit weights. Going forward, instead we shall consider the log-likelihoods L_1^{tb} instead to be gaussian-distributed variables to be combined by using their correlations, signals-to-noise, and average values – given, respectively, by $\mathcal{C}$, $\vec{s}$, and $\vec{\ell}$.

Total S/N for correlated variables

In the general case, in which $\mathcal{C}$ is non-diagonal and the weight vector $\vec{e}$ is arbitrary, the total S/N $s_{\rm tot}$ is given by [5]

$$s_{\rm tot}^2 = \frac{(\vec{e}\cdot\vec{\ell})^2}{\sum_{i,j} e_i \ell_i \mathcal{C}_{ij} \ell_j e_j/(s_i s_j)} \ . \qquad \text{(IV.13)}$$

If we (erroneously) continue to combine L_1^{tb}- L_4^{tb} using unit weights, $s_{\rm tot}$ becomes

$$s_{\rm tot}^2 = \frac{(\sum_k \ell_k)^2}{\sum_{i,j} \ell_i \mathcal{C}_{ij} \ell_j/(s_i s_j)} = (0.617)^2 \ , \qquad \text{(IV.14)}$$

using the numerical values tabulated above – substantially degraded from the hypothetical no-correlation case where an S/N of 0.753 could be obtained.

For this correlated case, the weight vector with optimum discriminating power is given within a constant factor by

$$e_j^{\rm opt} = \frac{\sum_i (\vec{s}^{\,\dagger}\,\mathcal{C}^{-1})_i\, s_i/\ell_i}{\vec{s}^{\,\dagger}\,\mathcal{C}^{-1}\,\vec{s}}\ , \qquad \text{(IV.15)}$$

where $\vec{s}^{\,\dagger}$ is the transpose of $\vec{s}$. Using these optimum weights,

$$s_{\rm tot}^2 = \vec{s}^{\,\dagger}\,\mathcal{C}^{-1}\,\vec{s} = (0.619)^2\ . \qquad \text{(IV.16)}$$

The total S/N improved very slightly compared to using wrong (unit) weights.

Cross-check on S/N optimization

As a cross-check on $\vec{e}^{\,\rm opt}$ from Eq. (IV.15) and its implementation, one can ask a numerical extremization package [6] to solve for $\vec{e}^{\,\rm num}$ that maximizes $s_{\rm tot}$ from Eq. (IV.13). This check was done for a calculation that used all seven L^{tb}'s. The results for the optimum weights $\vec{e}^{\,\rm opt}_{\rm analytic}$ and $\vec{e}^{\,\rm opt}_{\rm numerical}$ were

$$\vec{e}^{\,\rm opt}_{\rm analytic} = \begin{pmatrix} 0.784 \\ 0.517 \\ 0.118 \\ 0.837 \\ 0.590 \\ 0.466 \\ 0.096 \end{pmatrix} \ ;\ \vec{e}^{\,\rm opt}_{\rm numerical} = \begin{pmatrix} 0.786 \\ 0.518 \\ 0.155 \\ 0.821 \\ 0.527 \\ 0.448 \\ 0.108 \end{pmatrix}\ .$$

These weights are somewhat different, particularly for $y_{\rm rms}$, the $5^{\rm th}$ element. However, the combined S/N ($s_{\rm tot}$) for the two optimizations had the same value 0.654. Evidently the optimum is broad.

For reasons that shortly will become apparent, most studies employed a numerical rather than than an analytic optimization.

Correlation of combined likelihood with top mass

When the combined top-*vs.*-background likelihood L^{tb} is a weighted sum of the L_i^{tb}'s, using the weight vector $\vec{e}$ as in Eq. (IV.10), its normalized covariance $\mathcal{C}_{mL}$ with the reconstructed top mass $m_{\rm MTL4J}$ is

$$\mathcal{C}_{mL} = \frac{\sum_k e_k \mathcal{C}_{mk} \ell_k / s_k}{\sqrt{\sum_{i,j} e_i \ell_i \mathcal{C}_{ij} \ell_j e_j/(s_i s_j)}}\ . \qquad \text{(IV.17)}$$

For example, using the analytic and numerical optimizations of $\vec{e}$ for all seven variables that was described in the previous section, $\mathcal{C}_{mL}$ was equal to 0.189 and 0.185, respectively. This is too large a top mass correlation to satisfy the goal of this work.

Combined optimization of S/N and top mass correlation

Given that a numerical approach to optimizing the weight vector $\vec{e}$ proved feasible, it made sense to expand this approach by searching numerically for the weight vector that maximizes the S/N from Eq. (IV.13), subject to the constraint that $\mathcal{C}_{mL}$, the correlation of L^{tb} with $m_{\rm MTL4J}$ from Eq. (IV.17), vanishes. Operationally this was done by extremizing $(s_{\rm tot}^2 + H\,\mathcal{C}_{mL}^2)$, where H was a constant of order 10^3. For the same seven-variable example discussed above, the combined optimization yielded

$$\vec{e}^{\,\rm opt}_{\rm combined} = \begin{pmatrix} 0.815 \\ 0.629 \\ 0.521 \\ 1.108 \\ 0.346 \\ -0.014 \\ -0.104 \end{pmatrix}\ .$$

As required by the constraint, this radically different weight vector did yield $\mathcal{C}_{mL} = 0.000$, but the combined S/N was degraded to 0.625 from the unconstrained value of 0.654. In the new weight vector, variables 6 and 7, H_{T2} and $K_{T\rm min}$, entered only weakly and with negative coefficients. Evidently, in the constrained case, their correlation with $m_{\rm MTL4J}$ was too severe to allow them to participate effectively in separating signal from background.

Using fewer variables for optimization

Dropping variables 6 and 7 altogether, with $\mathcal{C}_{mL}$ again constrained to vanish, the combined S/N edged downward from 0.625 to 0.621. In this constrained optimization, variable 5 ($y_{\rm rms}$) received less than $\frac{1}{3}$ of the weight of any other variable.

Dropping each of the remaining 5 variables, it was confirmed that dropping $y_{\rm rms}$ yielded the smallest degradation: S/N edged further downward from 0.621 to 0.619 (the same value displayed in Eq. (IV.16)).

To summarize, with L^{tb} constrained to be uncorrelated with $m_{\rm MTL4J}$, the S/N degraded by 1%, from 0.625 to 0.619, as the number of variables was reduced from seven to the first four.

The choice to proceed with the four variables v_1-v_4 was made not only for the sake of simplicity. For mutual consistency and a cross-check, we sought also to use the same set of variables for the likelihood-based (LB) and neural network-based (NN) analyses. As seen above, for the LB analysis, it was possible using constrained numerical optimization to null the correlation of the overall likelihood L^{tb} with top mass, even when individual variables like v_6 and v_7 were significantly so correlated. For the NN analysis, no such technique was available to DØ at that time.

Finally, the likelihood discriminant D_{LB} was defined by

$$\mathcal{D}_{LB} = \frac{1}{1 + P\exp\left(-L^{tb}\right)} \ , \qquad \text{(IV.18)}$$

where P is a background/top prior set equal to 1.25 for Run 1 analysis (based on the expected mixture of background and top in that sample). D_{LB} is unity for top, zero for background, and more generally intended to be proportional to $P(\text{top})/(P(\text{top}) + P(\text{background}))$.

Scaling the weight vector

Handling two classes of background

As mentioned in Section II, the experimental background consisted of 82% W+jets and 18% fake isolated leptons ("QCD"); all of the figures in this note use that mixture. Correspondingly, all of the log-likelihoods L_i^{tb} described up to this point were "trained" on that mixture.

However, all of the numbers tabulated in the text up to this point, for example Eqs. (IV.4), (IV.6), (IV.8), (IV.12), *etc.*, as well as quoted values for $s_{\rm tot}$ and $\vec{e}$, result from applying mixture-trained L_i^{tb}'s to samples comprised of top and of background consisting only of 100% W+jets. Therefore it remains to describe the incorporation of QCD events into the scheme. We start with a short history.

As mentioned in Section II's discussion of missing E_T, originally this analysis set the more ambitious goal of distinguishing among all three event classes. Correspondingly, two sets of L^{tb}'s – $L^{tb}_{W+\rm jets}$ and $L^{tb}_{\rm QCD}$ – were formed using separate training samples for each background class. Next, two sets of $\vec{e}$'s were determined, optimizing S/N for separation of W+jets from top (using $L^{tb}_{W+\rm jets}$) and for QCD (using $L^{tb}_{\rm QCD}$. Enforcing

Text above figure.

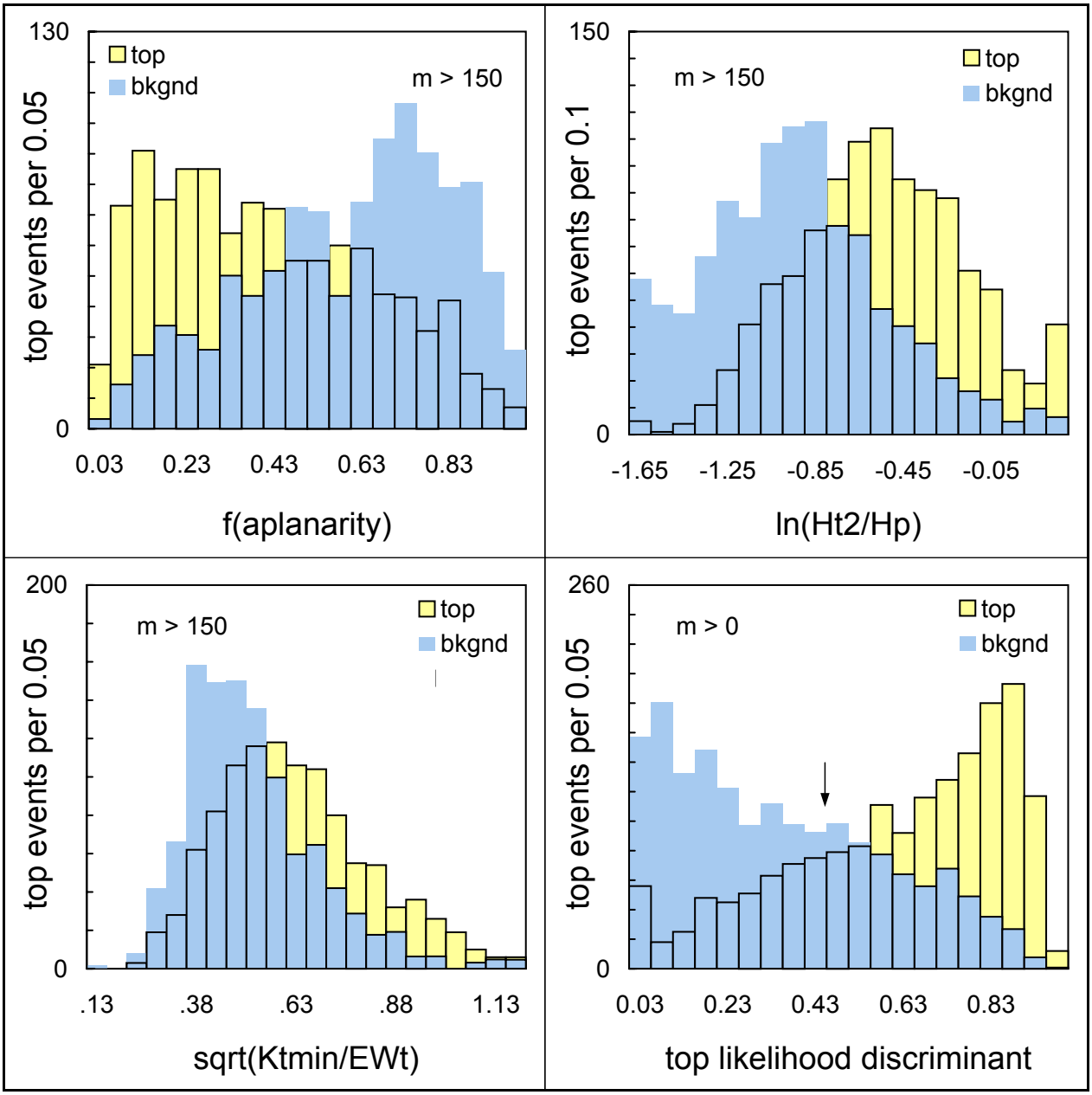

Figure caption.

Text below figure.

REFERENCES

[1] Top ℓ+jets mass analysis note
[2] Top ℓ+jets mass PRD
[3] Top discovery PRD
[4] Top ℓ+jets mass PRL
[5] These and other formulæ in Section V were derived by minimizing

$$\chi^2 = (\vec{x}^* - \vec{x}_i)^\dagger \, \mathcal{H} \, (\vec{x}^* - \vec{x}_i) \, ,$$

where $\vec{x}_i$ is the i^{th} measurement of a vector of observables, $\vec{x}^*$ is the fit vector, and $\mathcal{H}$ is the inverse covariance matrix. [6] This work used

EXCEL FINDER, but a more primitive package such as MINUIT presumably would suffice.